МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Институт математики, физики и информационных технологий
(наименование института полностью)

Кафедра «Прикладная математика и информатика»
(наименование)

02.03.03 Математическое обеспечение и администрирование информационных систем
(код и наименование направления подготовки, специальности)

Технология программирования
(направленность (профиль)/специализация

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

на тему: «Прогнозирование распространения вируса COVID-19 на основе статистического анализа данных»

Студент К.В. Щеглова
(И.О. Фамилия) (личная подпись)

Руководитель к.ф.-м.н., О.В. Лелонд
(ученая степень, звание, И.О. Фамилия)

Консультант М.В. Дайнеко
(ученая степень, звание, И.О. Фамилия)

Тольятти 2020

АННОТАЦИЯ

Тема бакалаврской работы: «Прогнозирование распространения вируса COVID-19 на основе статистического анализа данных».

В данной бакалаврской работе исследуется способы прогнозирования распространения вирусной инфекции COVID-19 на основе анализа статистических данных.

В работе предложен алгоритм составления модели прогнозирования ежедневного числа умерших от вирусной инфекции COVID-19 на основе статистических данных Всемирной организации здравоохранения. Алгоритм заключается в аппроксимации данных о смертности с помощью логистической функции путем определения набора ее параметров отдельно для каждой страны. Полученные функции используется для прогнозирования изменения числа смертей от инфекции COVID-19 в различных странах.

Структура бакалаврской работы представлена введением, тремя разделами, заключением, списком литературы.

Во введении описывается актуальность проводимого исследования, дается краткая характеристика проделанной работы.

В первом разделе проводится анализ перспектив развития алгоритмов для прогнозирования глобальных вирусных инфекций.

Во втором разделе описывается математический аппарат предложенного алгоритма прогнозирования вирусной инфекции COVID-19, дается описание источника данных по распространению, приводятся результаты прогнозирования, оценивается их точность.

В третьем разделе дается описание программного обеспечения для реализации предложенных подходов по прогнозированию.

В заключении представлены выводы по проделанной работе.

В работе использовано 2 таблицы, 31 рисунков, список литературы содержит 20 литературных источников. Объем бакалаврской работы составляет 50 страниц.

ABSTRACT

The theme of the bachelor's work: "Prediction of COVID-19 virus spread based on statistical analysis of data".

This bachelor's work explores ways to predict the spread of COVID-19 virus infection based on statistical analysis.

In work the algorithm of drawing up of model of forecasting of daily number of dead from virus infection COVID-19 on the basis of statistics of the World organization of public health services is offered. The algorithm consists in approximating the mortality data using a logistics function by defining a set of its parameters separately for each country. The resulting functions are used to predict changes in the number of deaths from COVID-19 infection in different countries.

The structure of the bachelor's work is represented by an introduction, three chapters, an opinion and a list of literature.

The introduction describes the relevance of the ongoing research and gives a brief description of the work done.

The first chapter analyzes the prospects for developing algorithms for predicting global viral infections.

The second chapter describes the mathematical apparatus of the proposed algorithm for the prediction of virus infection COVID-19, gives a description of the source of distribution data, gives the results of the prediction, and assesses their accuracy.

The third chapter describes the software to implement the proposed prediction approaches.

Finally, conclusions are presented on the work done.

In the work 2 tables, 31 figures are used, the list of literature contains 20 literature sources. The volume of bachelor's work is 50 pages.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ 5

1 АНАЛИЗ ПЕРСПЕКТИВ РАЗВИТИЯ АЛГОРИТМОВ СНИЖЕНИЯ ПРИЗНАКОВОГО ПРОСТРАНСТВА 7

2 РАЗРАБОТКА АЛГОРИТМА СНИЖЕНИЯ ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ ВРЕМЕННЫХ РЯДОВ 9

2.1 Математический аппарат 9

2.2 Источник данных для анализа 10

2.3 Источник данных для анализа 12

2.3 Оценка точности прогнозов, построенных по предложенной модели 23

3 ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ПРЕДЛОЖЕННЫХ РЕШЕНИЙ 25

3.1 Особенности разработанного программного обеспечения 25

3.2 Описание программного кода 25

ЗАКЛЮЧЕНИЕ 44

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ 46

ВВЕДЕНИЕ

Всемирной организация здравоохранения в этом году была объявлена чрезвычайная ситуация связанная с распространением коронавирусной инфекции COVID-19.

Коронавирусная инфекция COVID-19 приводит к необратимым изменениям в тканях легких человека, а при возникновении осложнений может привести к смерти. На момент написание данной бакалаврской работы в мире уже зарегистрировано более 3 миллионов случаев заражения коронавирусом. Помимо биологических угроз вирус COVID-19 приводит и к негативным экономическим последствиям. Это объясняется тем, что во многих странах мира для обеспечения безопасности населения вводится карантин или меры по ограничению передвижения людей. Многих сотрудников переводят на дистанционную форму работы. В связи с этим останавливаются многие производства, закрывается большое количество малых и средних предприятий. Это приводит к снижению роста экономик большинства развивающихся стран.

Для минимизации вреда силы многих исследователей были брошены на изучение различных явлений, связанных с COVID-19. Современные работы в области анализа данных направленны на исследования способов ранней диагностики COVID-19, определение групп риска людей на основе анализа статистики, прогнозирования тяжести протекания заболевания, а также определения дат снижения прироста числа заболевших. Таким образом, любые исследования связанные с COVID-19 в настоящее время являются актуальными.

В данной бакалаврской работе исследуется способы прогнозирования распространения вирусной инфекции COVID-19

Целью работы является разработка алгоритма прогнозирования распространения вирусной инфекции COVID-19 на основе статистического анализа данных.

В работе предложен алгоритм прогнозирования распространения вирусной инфекции COVID-19, который заключается в выполнении следующих шагов. Загрузка статистических данных Всемирной организации здравоохранения о ежедневном количестве зарегистрированных смертей от вируса (отдельно по каждой стране). Затем, аппроксимация полученных данных логистической функцией путем подбора параметров функции методом наименьших квадратов. Так как динамика распространения COVID-19 внутри каждой страны уникальна, то и параметры логистической функции для каждой страны свои собственные. Затем полученные функции используются для оценки изменения количества смертей по каждой стране в ближайшее время.

В рамках данных исследования было разработано программное обеспечение, позволяющее получать актуальные данные о количестве зарегистрированных смертей по вирусу COVID-19 и пересчитывать, с учетом новых данных, параметры логистических функций. Результат аппроксимации исходных данных представляется графически в виде графиков временных рядов: количество зарегистрированных смертей в зависимости от даты и прирост количества зарегистрированных смертей в зависимости от даты.

1 АНАЛИЗ ПЕРСПЕКТИВ РАЗВИТИЯ МЕТОДОВ ДЛЯ ПРОГНОЗИРОВАНИЯ РАСПРОСТРАНЕНИЯ ИНФЕКЦИЙ МИРОВОГО МАСШТАБА

Для сдерживания распространения вируса COVID-19 государства всех стран вынуждены принимать меры, оказывающие негативное влияние на экономику.

Одной популярных мер является перепрофилирование больниц для приема многочисленных больных COVID-19. При этом перестает оказываться либо снижается помощь больным с другими профильными заболеваниями. Само по себе перепрофилирование больниц связано со значительными денежными затратами из-за необходимости закупки дополнительного оборудования, а также оснащения и наращивания койко-мест.

Другой мерой по сдерживанию COVID-19 является введения ограничительных мер, связанных со свободой передвижения граждан по улицам и в других общественных местах. Для слежения за соблюдением введенных ограничительных мер требуется задействовать перевод большого количества сотрудников полиции с других работ на патрулирование.

Из-за необходимости снижения скорости распространения COVID-19 вводятся дополнительные нерабочие, но оплачиваемые дни. В этом случае некоторые предприятия вынуждены сокращать численность штата, т.к. не могут выплачивать зарплаты при отсутствии выручки от остановленной деятельности предприятий. В таких случаях государство вынуждено оказывать финансовую поддержку предприятиям и выплачивать пособия гражданам, оставшимся без работы.

Также для сдерживания распространения COVID-19 государство вынуждено тратить средства на проведение научных исследований инфекции, а также разработку тестов и вакцин.

Негативные экономические последствия от введения данных мер можно снизить путем правильного выбора времени по их внедрению. Но для того, чтобы планировать введение и снятие описанных выше мер необходимо владеть методами прогнозирования распространения COVID-19.

Таким образом, актуальной проблемой остается разработка программных средств и алгоритмов прогнозирования распространения инфекций в масштабах как отдельных стран, так и мира в целом.

Целью работы является разработка алгоритма прогнозирования распространения вирусной инфекции COVID-19 на основе статистического анализа данных.

Поставленная цель будет достигнута за счет решения следующих задач:

1. Анализ перспектив развития методов для прогнозирования распространения инфекций мирового масштаба.

2. Разработка алгоритма прогнозирования распространения COVID-19 (смертности от инфекции).

3. Разработка программного обеспечения, реализующего получение статистических данных и составление прогноза по распространению COVID-19 основе предложенного алгоритма.

Выводы по разделу

Актуальность темы обусловлена заявлением Всемирной организацией здравоохранения о признании COVID-19 в 2020 году пандемией. Поэтому все исследования, направленные на изучение особенностей распространения COVID-19 являются в настоящее время актуальными.

По результатам материалов, представленных в данной главе, были сформулированы цели и задачи исследования.

2 РАЗРАБОТКА АЛГОРИТМА ПРОГНОЗИРОВАНИЯ РАСПРОСТРАНЕНИЯ COVID-19

2.1 Математическая модель распространения COVID-19

Из открытых научных статей, посвященных развитию распространения различных вирусов известно, что в начале распространения любого вируса количество смертей незначительно. Это объясняется малой распространённостью вируса внутри популяции. В зараженных особях вирус внутри организма еще не размножился до той степени, чтобы обеспечить ярко выраженные симптомы, угрожающие жизни [1-3].

Затем скорость распространения вируса внутри популяции увеличивается. Рост скорости распространения вируса происходит до определенного момента времени t_c, который называется «пиком заболеваемости». После данного момента времени скорость распространения вируса начинает уменьшаться. Преодоление пика заболеваемости связано либо с выработкой группового иммунитета, либо с применение вакцинации, либо с принятием удачного комплекса мер по сдерживанию вируса [4-8].

Поддержка снижения скорости распространения вируса в дальнейшем приводит к тому, что количество новых случаев заражение становится незначительным. Достижение данного момента времени t_p, называется выходом на «плато заболеваемости» [9-11].

Так как наибольший вред стране наносят не носители COVID-19, а человеческие жертвы, связанные с вирусом, то целесообразен анализ не динамики заболеваемости, а количества смертей от вируса. С учетом данного факта и ключевых временных точек (пик заболеваемости t_c и выход t_p на плато заболеваемости) теоретические кривые смертности от COVID-19 будут выглядеть так, как это показано на рисунке 2.1.

Так как теоретическая кривая распространения вируса по форме напоминает логистическую функцию, то в исследовании предполагается, что

статистические данные о количестве смертей от COVID-19 можно аппроксимировать логистической функцией $f_1(t)$ (2.1).

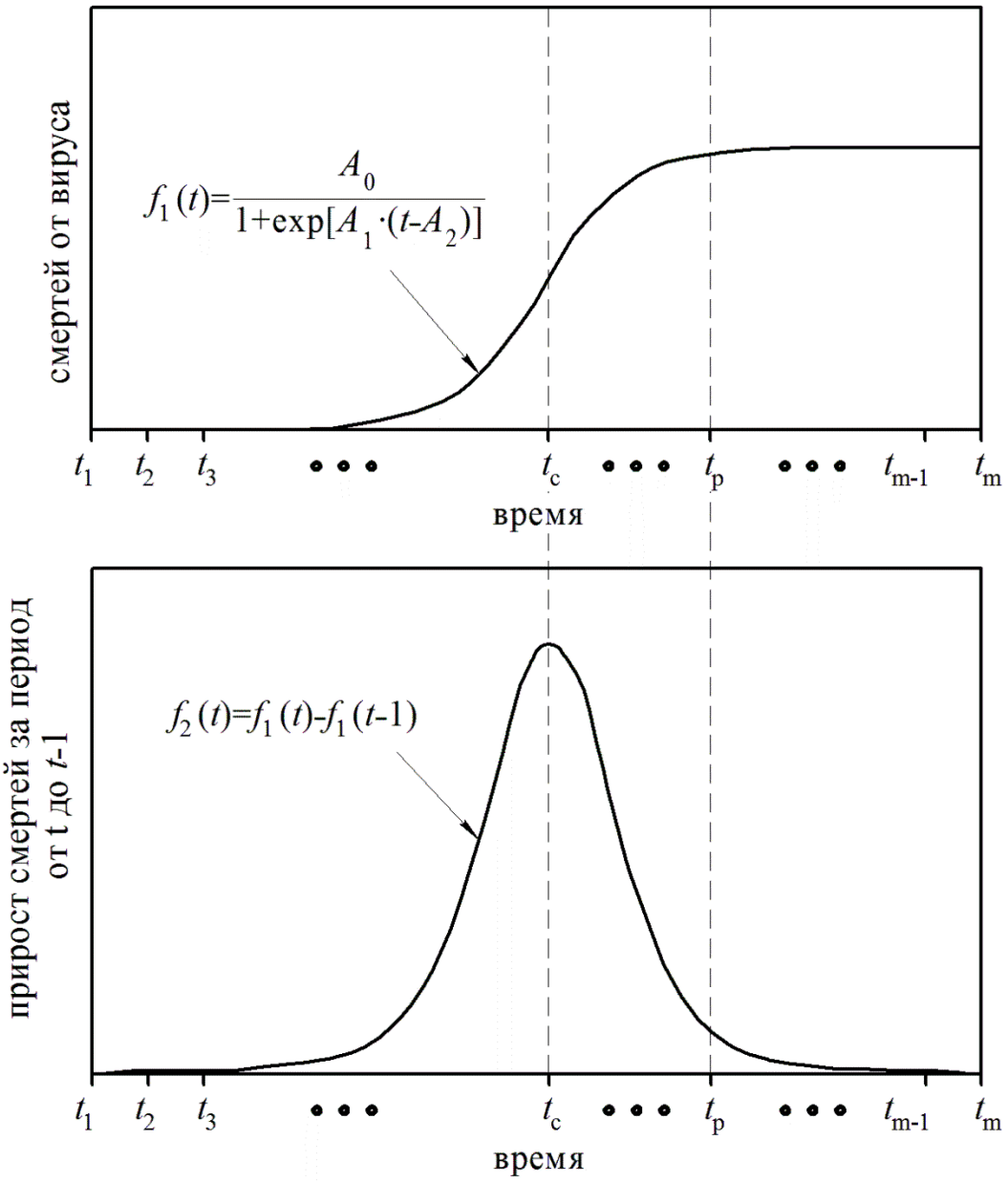

Рисунок 2.1 – Модель распространения вируса в течение времени

$$f_1(t) = \frac{A_0}{1+e^{A_1(t-A_2)}}, \tag{2.1}$$

где A_0, A_1, A_2 – параметры логистической функции.

В каждой стране принимается разный набор мер по противодействию распространения COVID-19. Например, в одних странах вводится карантин, в

других странах самоизоляция является рекомендательной мерой. В одних странах маски выдаются населению бесплатно, в других странах население вынуждено покупать одноразовые маски за свой счет. Поэтому коэффициенты A_0, A_1, A_2 логистической функции описывающей данные о смертях от COVID-19 у каждой страны будут свои.

В исследовании предполагается, что если для заданной страны удастся подобрать такие параметры A_0, A_1, A_2, при которых логистическая функция будет достаточно точно описывать фактические данные, то эту функцию можно будет использовать для прогнозирования смертности от COVID-19 (в будущем).

Определение параметров A_0, A_1, A_2 возможно методом наименьших квадратов. В этом случае задачу аппроксимации фактических данных $fact(t)$ функцией $f_1(t)$ можно представить формулой (2.2).

$$\begin{cases} f_1(t, A_0, A_1, A_2) = \dfrac{A_0}{1 + e^{A_1 (t - A_2)}} \\ \sum\limits_{t \in [t_1; t_n]} (f_1(t, A_0, A_1, A_2) - fact(t))^2 \to \min \end{cases}, \quad (2.2)$$

где t – номер дня с начала года, $[t_1; t_n]$ – диапазон времени, внутри которого известны фактические значения по смертности от COVID-19, A_0, A_1, A_2 – параметры логистической функции, которые подбираются при решении задачи оптимизации.

Таким образом, алгоритм прогнозирования смертности от COVID-19 включает в себя следующие шаги.

1. Получение фактических данных (временной ряд) Всемирной организации здравоохранения о смертности от COVID-19.

2. Поиск параметров A_0, A_1, A_2 логистической функции, обеспечивающих наиболее точно описание фактических данных.

3. Использование полученной функции для прогнозирования смертности от COVID-19 в будущем (на временной интервал, где фактические данные пока не известны).

2.2 Источник данных для анализа

В качестве источника информации по ежедневному количеству смертей от COVID-19 был выбран репозиторий данных университета Джона Хопкинса (Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University), который расположен в сервисе github по адресу: https://github.com/CSSEGISandData/COVID-19.

В качестве основного источника данных по COVID-19 в данном репозитории указана информация с сайта Всемирной организации здравоохранения (World Health Organization), расположенного по адресу https://www.who.int/.

Другими источниками данных репозитория являются: BNO News, National Health Commission of the People's Republic of China (NHC), China CDC (CCDC), Hong Kong Department of Health, Macau Government, Government of Canada, Italy Ministry of Health, French Government, Washington State Department of Health и др.

Данный репозиторий создавался и поддерживается ежедневно в актуальном состоянии для визуализации данных по COVID-19 в виде сервиса «Visual Dashboard by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)». Данные по распространению COVID-19 хранятся в репозитории в папке csse_covid_19_data, внутри которой находится множество файлов в формате csv.

Так как нас интересуют только данные по смертям от COVID-19, мы будем использовать данные, расположенные в файле time_series_covid19_deaths_global.csv. В нем хранятся данные о ежедневном количестве смертей от COVID-19 для различных регионов 185 стран мира.

На основе этих данные построены все последующие графики и найдены описанные ниже аппроксимирующие функции.

2.3 Результат анализа статистических данных на 4 апреля

Для тестирования предложенного подхода, а также для сравнения результатов аппроксимации и прогнозирования были выбраны контрольные временные точки – статистические данные о смертях от COVID-19, актуальные на 04.04.2020 и на 29.04.2020.

По результатам анализа статистических данных для каждой страны и мира в целом были построены аппроксимирующие функции, описывающие распространение COVID-19 в течение времени.

Для каждой страны, а также для мира в целом, было построено по два графика: количество смертей от COVID-19 всего в зависимости от даты и ежедневный прирост смертей от COVID-19 в зависимости от даты. Здесь и далее на графиках крестиками обозначены фактические данные, имеющиеся в распоряжении Всемирной организации здравоохранения на текущую дату, а синей линией обозначен график логистической функции, аппроксимирующей имеющиеся данные. На временном интервале, где отсутствуют фактические данные, аппроксимирующая логистическая функция играет роль модели прогнозирования.

Данные графиков изменения ежедневного прироста смертей получают путем вычитания количества смертей на рассматриваемую дату и количества смертей за день до рассматриваемой даты.

Вообще логистические аппроксимирующие функции были получены для 185 стран, но в работе для краткости будут приведены графики распространения COVID-19 для всего мира в целом (рисунок 2.2), для Китая (рисунок 2.3), а также для США (рисунок 2.4).

Дальнейший анализ полученных аппроксимирующих зависимостей позволит составить прогнозы об изменении эпидемиологической ситуации по COVID-19 в будущем.

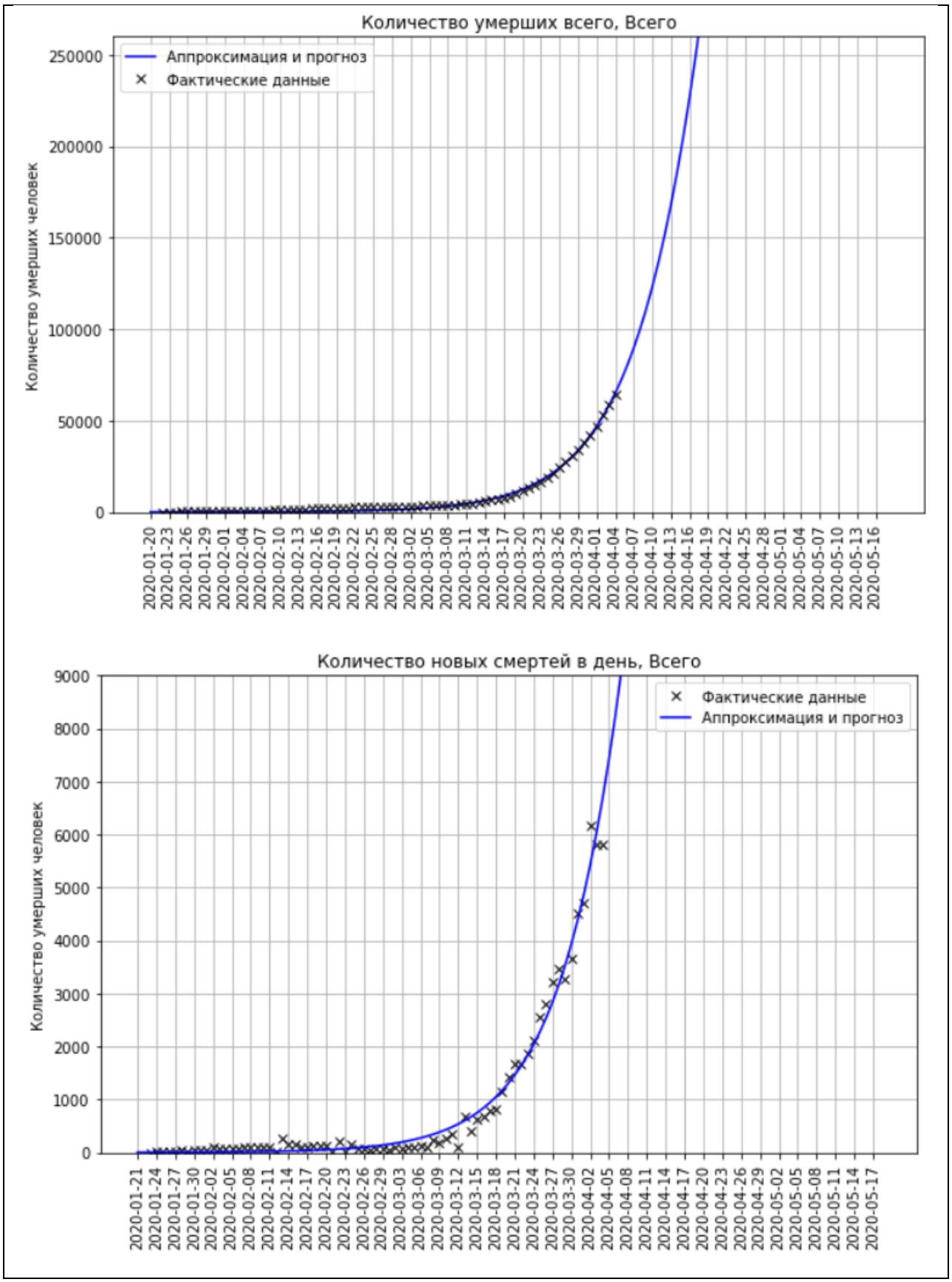

Рисунок 2.2 – Результаты анализа данных по смертности по всем странам

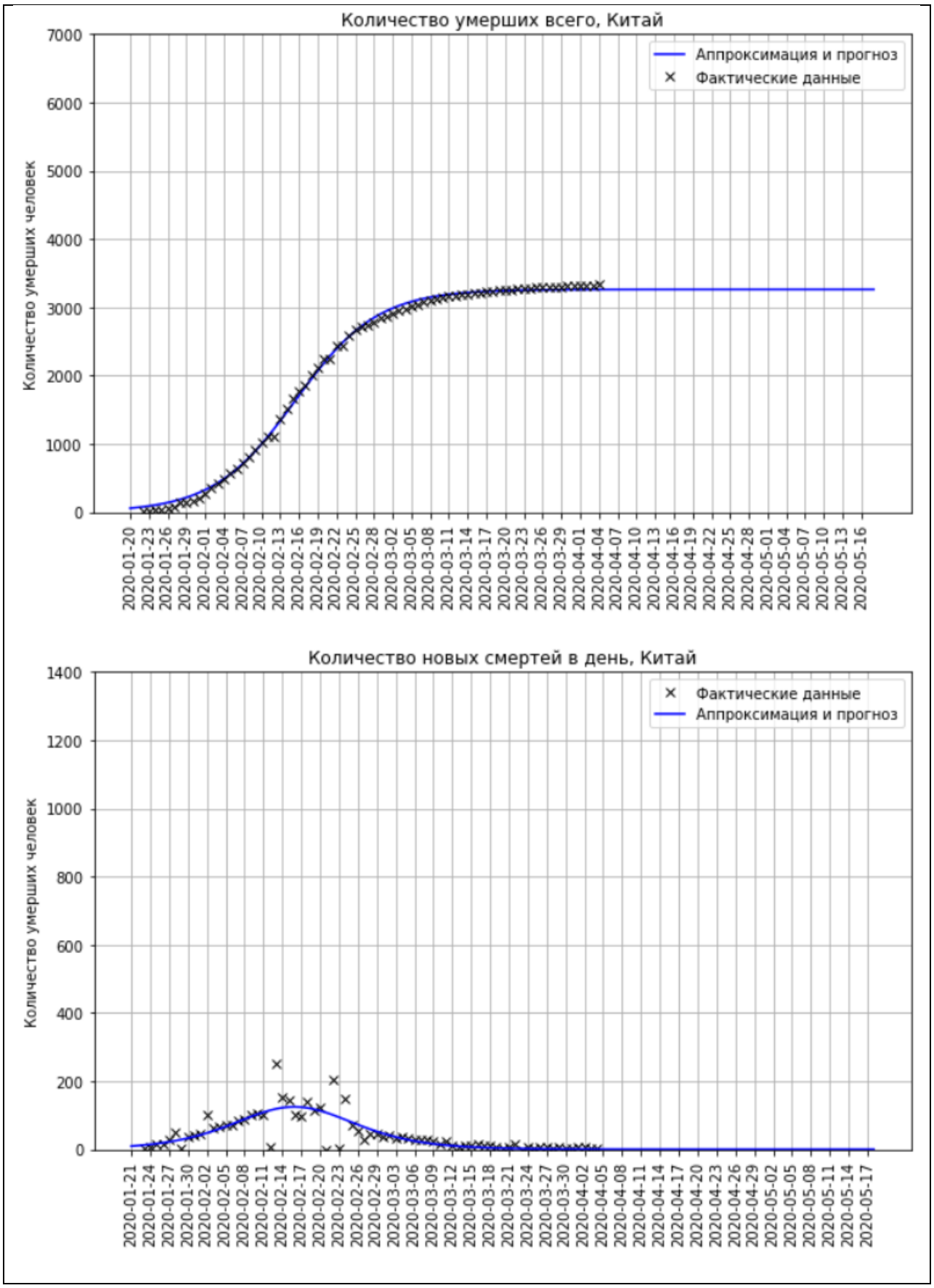

Рисунок 2.3 – Результаты анализа данных по смертности для Китая

В ходе анализа статистических данных по смертности от COVID-19, актуальных на 4 апреля 2020, можно сделать следующие прогнозы.

В целом по миру (рисунок 2.2), судя по графику логистической функции, в течение 7 дней (на временном интервале 04.04.2020-11.04.2020) сохранится рост числа смертей от COVID-19. О том, что в течение ближайшей недели не будет достигнут пик смертей, говорит отсутствие максимума на нижнем графике рисунка 2.2. Если говорить о численных оценках, то конец интервала прогнозирования (11.04.2020), составляющего неделю, общее количество смертей от COVID-19 достигнет значения 125000.

В Китае (рисунок 2.3), судя по графику логистической функции, в ближайшую неделю (04.04.2020-11.04.2020) будет отсутствовать значительный рост смертей от COVID-19. Также по нижнему графику рисунка 2.3 можно отметить, что пик по показателю прироста смертей пройден и, судя по линии логистической функции, он был ориентировочно 16 февраля 2020 года. Если говорить о численных оценках, то на конец интервала прогнозирования (11.04.2020) общее количество смертей в Китае от COVID-19 будет составлять 3200-3400.

В США (рисунок 2.4), судя по верхнему графику в ближайшую неделю (04.04.2020-11.04.2020) сохранится рост числа смертей от COVID-19. Если говорить о численных оценках, то на конец интервала прогнозирования (11.04.2020) общее количество смертей в США от COVID-19 достигнет значения 19500.

Отдельно стоит отметить, что хотя все графики логистических функций и нарисованы до середины мая 2020 года, но прогнозы дальше 7 дней от последнего фактического значения обладают малой точностью.

Достоверность поставленных прогнозов будет проверяться в следующем параграфе работы на основе актуальных на 29 апреля данных.

Рассчитанные коэффициенты логистической функции для разных стран представлены в таблице 2.1.

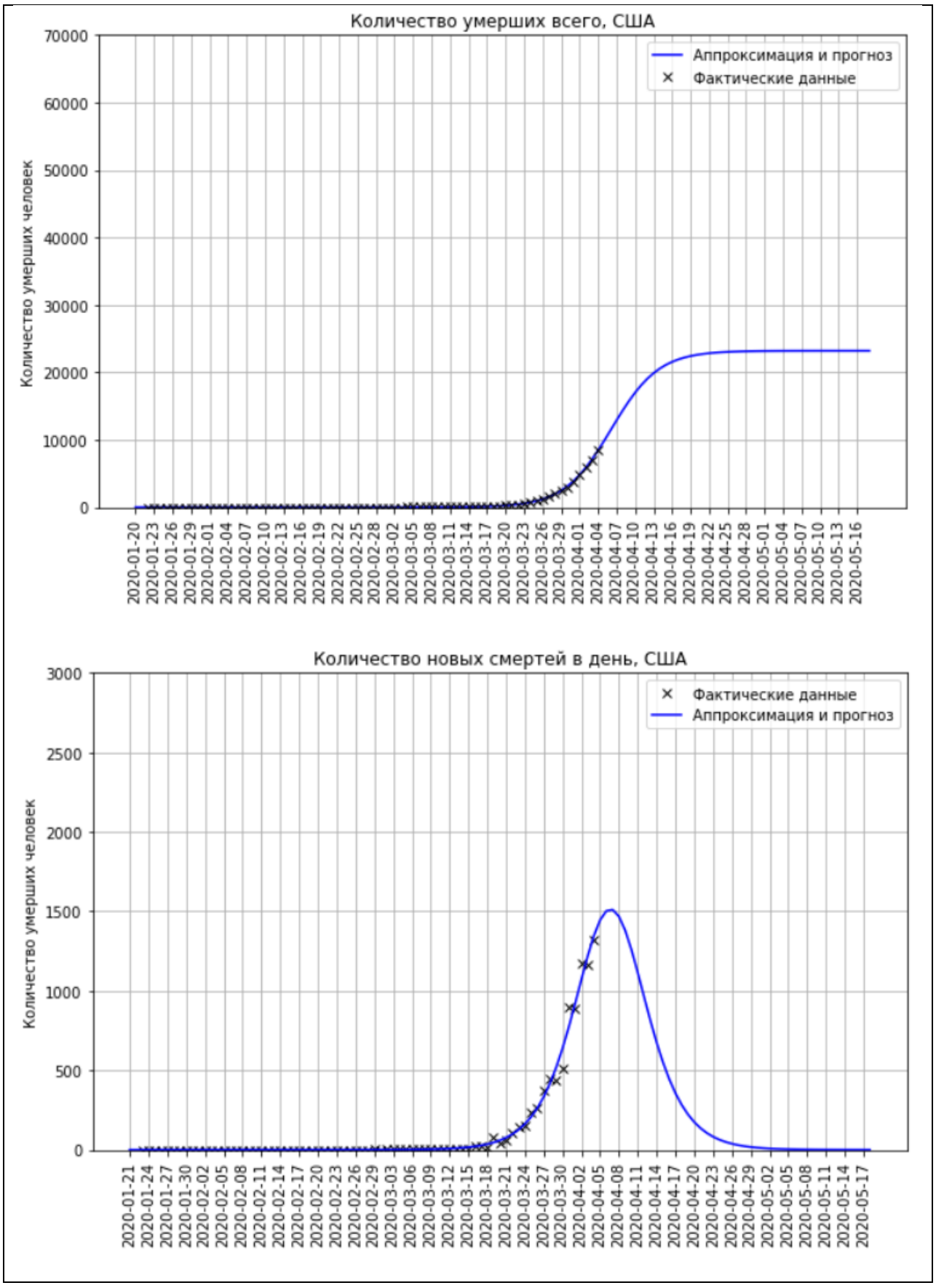

Рисунок 2.4 – Результаты анализа данных по смертности для США

Таблица 2.1 – Результаты аппроксимации количества умерших в зависимости от номера дня с начала 2020 для некоторых стран (по данным на 04.04.2020)

Страна	Коэффициенты логистической функции		
	k_0	k_1	k_2
Китай	3.26233659×10^{3}	$-1.52735843 \times 10^{-1}$	4.64021635×10^{1}
Франция	1.23767194×10^{5}	$-1.87382451 \times 10^{-1}$	1.09565046×10^{2}
Иран	3.78824474×10^{3}	$-1.69868666 \times 10^{-1}$	8.35861284×10^{1}
Италия	1.86528074×10^{4}	$-1.84865545 \times 10^{-1}$	8.71262970×10^{1}
Испания	1.46504905×10^{4}	$-2.58242894 \times 10^{-1}$	8.95396142×10^{1}
Англия	3.26010763×10^{4}	$-2.32143263 \times 10^{-1}$	1.03047131×10^{2}
Америка	2.31954870×10^{4}	$-2.61431171 \times 10^{-1}$	9.71480421×10^{1}
Бельгия	2.14409433×10^{3}	$-2.93087295 \times 10^{-1}$	9.35608410×10^{1}
Германия	2.62248555×10^{3}	$-2.64900884 \times 10^{-1}$	9.42291032×10^{1}
Нидерланды	2.45282017×10^{3}	$-2.50274259 \times 10^{-1}$	9.22367244×10^{1}
Швейцария	1.07895633×10^{3}	$-2.23567306 \times 10^{-1}$	9.29824799×10^{1}
По всем странам	1.69176558×10^{6}	$-1.12923891 \times 10^{-1}$	1.23516775×10^{2}

2.4 Результат анализа статистических данных на 29 апреля

Аналогичный анализ проведен по данным, актуальным на 29.04.2020. Полученные графики показаны на рисунках 2.5, 2.6 и 2.7.

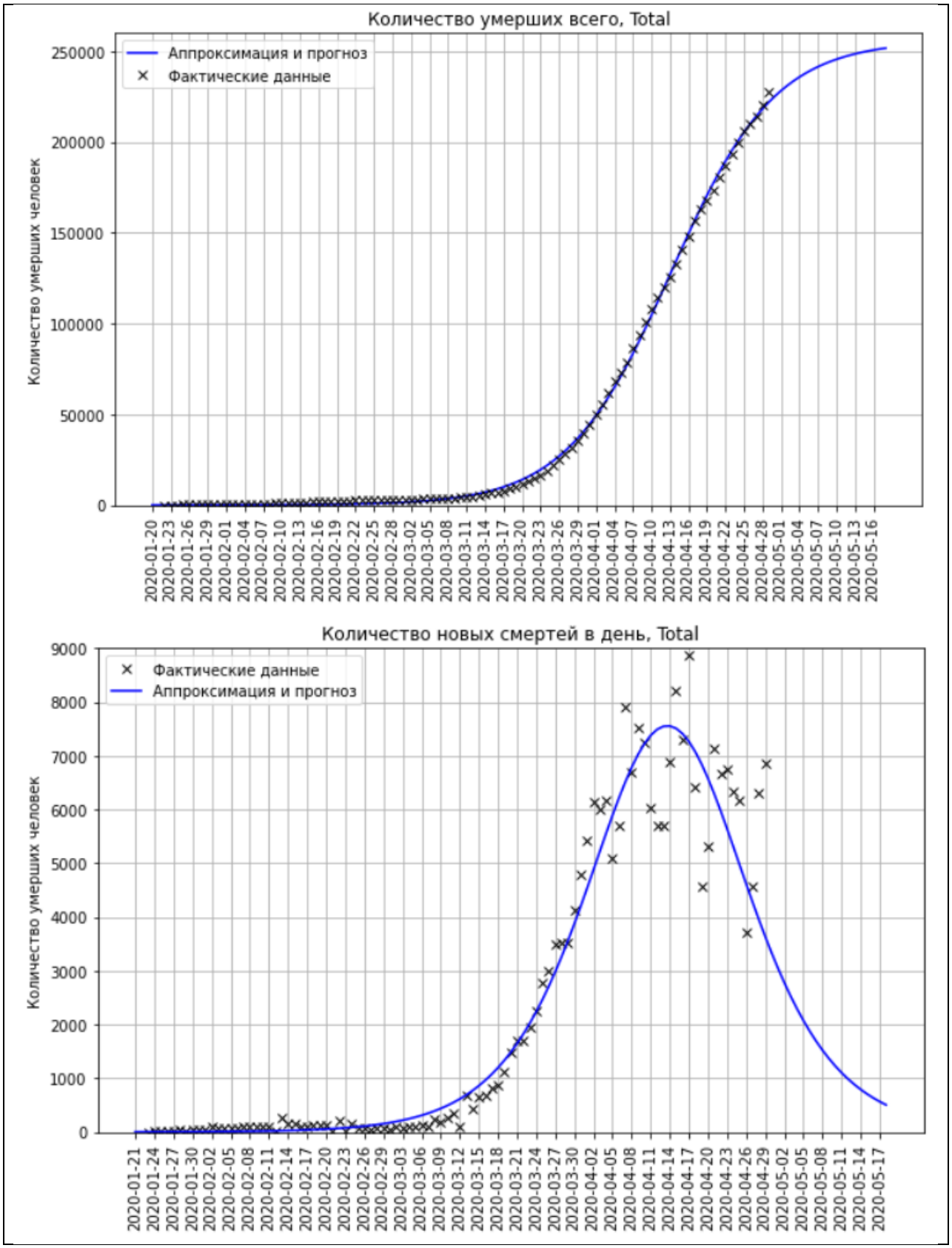

Рисунок 2.5 – Результаты анализа данных по смертности по всем странам

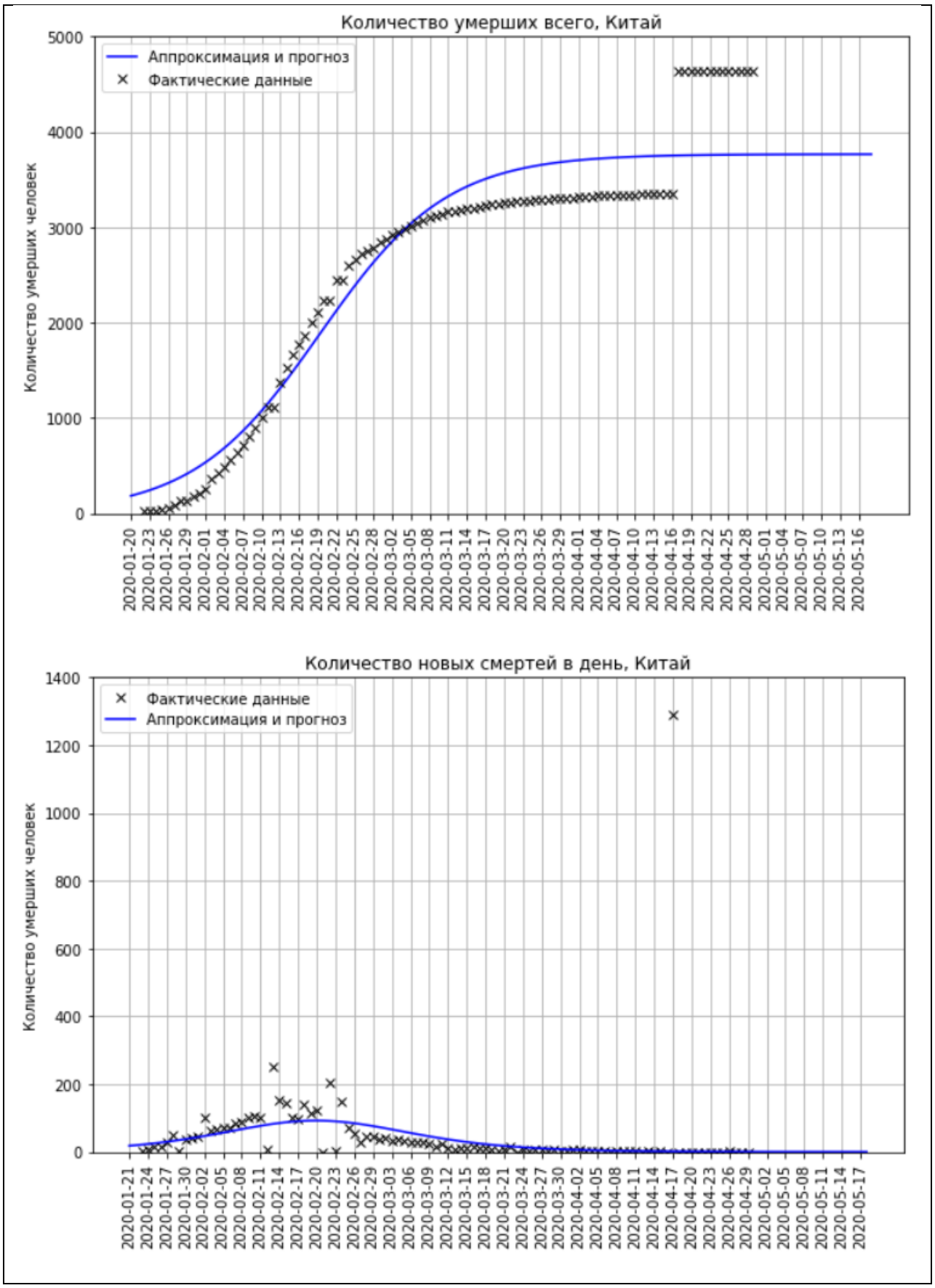

Рисунок 2.6 – Результаты анализа данных по смертности для Китая

На рисунках 2.5-2.7 представлены обновленные статистические данные по смертям от COVID-19, а также графики логистических функций с пересчитанными значениями коэффициентов A_0, A_1, A_2. По представленным данным проверим точность прогнозов, описанных в предыдущем параграфе.

В целом по миру (рисунок 2.5) на 11.04.2020 прогнозировался рост количества смертей до уровня 125000 человек. В действительности количество смертей на указанную дату достигло уровня 110000 человек. Таким образом, точность прогноза составила – 87%.

В Китае (рисунок 2.6) на 11.04.2020 прогнозировалось сохранение общего числа смертей на уровне 3200-3400. Судя по фактическим данным, это прогноз оказался полностью верным.

В США (рисунок 2.6) на 11.04.2020 прогнозировался рост общего числа смертей до уровня 19500. В действительности количество смертей на указанную дату достигло уровня 22000 человек. Таким образом, точность прогноза составила – 89%

Полученные результаты подтверждают, что на интервале 7 дней точность прогнозов по количеству смертей от COVID-19 остается на уровне не ниже 85%. Однако такая точность сохраняется в том случае, если в собранных статистических данных нет ошибок. Например, если взглянуть на верхнюю часть графика, представленного на рисунке 2.6, то можно увидеть одномоментное резкое увеличение общего числа смертей в Китае на 14.04.2020. Это официально объясняется тем, что изначально статистика Китая была занижена, и начиная с 14.04.2020 публикуются уточнённые значения. Такие всплески данных приводят к существенному снижению точности прогнозов с использованием логистической функции.

Рассчитанные коэффициенты логистической функции для разных стран на основе новых статистических данных представлены в таблице 2.2.

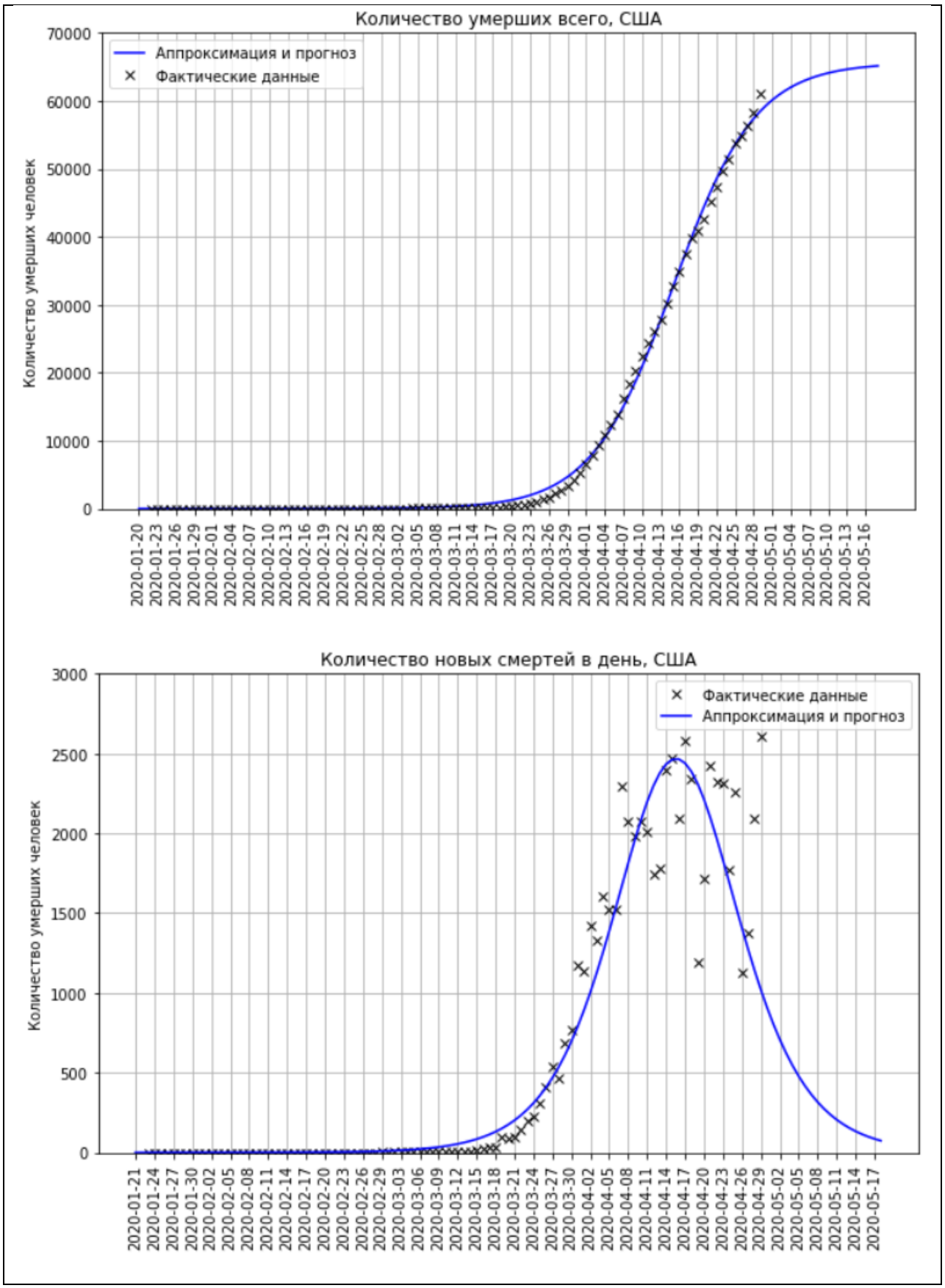

Рисунок 2.7 – Результаты анализа данных по смертности для США

Таблица 2.2 – Результаты аппроксимации количества умерших в зависимости от номера дня с начала 2020 для некоторых стран (по данным на 01.05.2020)

Страна	Коэффициенты логистической функции		
	k_0	k_1	k_2
Китай	3.76688014×10^{3}	$-9.78655882 \times 10^{-2}$	5.03859815×10^{1}
Франция	2.42174062×10^{4}	$-1.68845004 \times 10^{-1}$	1.00500161×10^{2}
Иран	5.94260018×10^{3}	$-1.09582808 \times 10^{-1}$	9.18491360×10^{1}
Италия	2.74904904×10^{4}	$-1.22233441 \times 10^{-1}$	9.36995148×10^{1}
Испания	2.34207638×10^{4}	$-1.53829749 \times 10^{-1}$	9.55484531×10^{1}
Англия	2.75321025×10^{4}	$-1.55738757 \times 10^{-1}$	1.04730725×10^{2}
Америка	6.56096836×10^{4}	$-1.50549185 \times 10^{-1}$	1.06074874×10^{2}
Бельгия	7.76369936×10^{3}	$-1.75459048 \times 10^{-1}$	1.04144848×10^{2}
Германия	6.99423450×10^{3}	$-1.46295600 \times 10^{-1}$	1.04813872×10^{2}
Нидерланды	4.85486442×10^{3}	$-1.40924318 \times 10^{-1}$	1.00824995×10^{2}
Швейцария	1.71943841×10^{3}	$-1.48901299 \times 10^{-1}$	9.90011594×10^{1}
По всем странам	2.55869179×10^{5}	$-1.18201737 \times 10^{-1}$	1.04124904×10^{2}

2.3 Оценка точности прогнозов, построенных по предложенной модели

Результаты тестирования предложенного подхода на реальных данных показали следующее.

- Точность прогнозирования изменения количества смертей от COVID-19 с использованием логистической функции позволяет достигнуть значения не менее 85% на интервале 1 недели.
- Скачкообразное изменение статистических данных, связанное с занижением информации о смертях (см. рисунок 2.6), приводит к резкому снижению точности прогнозирования.
- Для поддержания точности прогнозов на высоком уровне необходимо с ежедневным получением новых статистических данных производить пересчет коэффициентов A_0, A_1, A_2.

Выводы по разделу

При анализе статистических данных могут применяться различные методы [12-20].

Разработан алгоритм прогнозирования распространения вирусной инфекции COVID-19, который заключается в выполнении следующих шагов. Сначала осуществляется загрузка статистических данных Всемирной организации здравоохранения о ежедневном количестве зарегистрированных смертей от вируса (отдельно по каждой стране). Затем проводится аппроксимация полученных данных логистической функцией путем подбора параметров функции методом наименьших квадратов.

Предложенный подход позволяет предсказывать количество смертей от COVID-19 на временном промежутке до 1 недели с приемлемой точностью – не менее 85%

3 ПРОГРАММНАЯ РЕАЛИЗАЦИЯ АЛГОРИТМА ПРОГНОЗИРОВАНИЯ

3.1 Особенности разработанного программного обеспечения

В ходе выполнения бакалаврской работы на языке Python было разработано программное обеспечение для анализа статистических данных Всемирной организации здравоохранения по распространению вируса COVID-19.

Разработанное программное обеспечение обладает следующими особенностями:

- реализована возможность прогнозирования общего числа смертей от COVID-19 и динамику ежедневного прироста смертности за счет аппроксимации статистических данных логистической функции;
- в приложении реализована оптимизация параметров аппроксимирующей логистической функции методом наименьших квадратов;
- реализована визуализация данных путем построения графиков аппроксимирующих функций и фактических данных по изменению общего числа смертей, а также по ежедневному приросту смертей от COVID-19;
- помимо обычной временной шкалы в формате число-месяц-год для удобства восприятия данных в приложении также реализована альтернативная шкала – номер дня от начала года;
- при каждом запуске приложения осуществляется загрузка из репозитория актуальных на текущий момент времени статистических данных (в виде CSV файла) и обновление их с учетом прогнозов по распространению COVID-19;
- существует возможность получать прогноз по количеству смертей от COVID-19 для выбранной страны на выбранную дату;

- в приложении реализовано построение прогнозов распространения коронавируса для 185 стран: Афганистан, Албания, Алжир, Андорра, Ангола, Антигуа и Барбуда, Аргентина, Армения, Австралия, Австрия, Азербайджан, Багамские острова, Бахрейн, Бангладеш, Барбадос, Беларусь, Бельгия, Бенин, Бутан, Боливия, Босния и Герцеговина, Бразилия, Бруней, Болгария, Буркина-Фасо, Кабо-Верде, Камбоджа, Камерун, Канада, Центральноафриканская Республика, Чад, Чили, Китай, Колумбия, Конго (Браззавиль), Конго (Киншаса), Коста-Рика, Кот-д'Ивуар, Хорватия, Diamond Princess, Куба Кипр, Чехия, Дания, Джибути, Доминиканская Республика, Эквадор, Египет, Сальвадор, Экваториальная Гвинея, Эритрея, Эстония, Эсватини, Эфиопия, Фиджи, Финляндия, Франция, Габон, Гамбия, Грузия, Германия, Гана, Греция, Гватемала, Гвинея, Гайана, Гаити, Святой Престол, Гондурас, Венгрия, Исландия, Индия, Индонезия, Иран, Ирак, Ирландия, Израиль, Италия, Ямайка, Япония, Иордания, Казахстан, Кения, Корея, Юг, Кувейт, Кыргызстан, Латвия, Ливан , Либерия, Лихтенштейн, Литва, Люксембург, Мадагаскар, Малайзия, Мальдивы, Мальта, Мавритания, Маури Тиус, Мексика, Молдова, Монако, Монголия, Черногория, Марокко, Намибия, Непал, Нидерланды, Новая Зеландия, Никарагуа, Нигер, Нигерия, Северная Македония, Норвегия, Оман, Пакистан, Панама, Папуа-Новая Гвинея, Парагвай, Перу, Филиппины, Польша, Португалия, Катар, Румыния, Россия, Руанда, Сент-Люсия, Сент-Винсент и Гренадины, Сан-Марино, Саудовская Аравия, Сенегал, Сербия, Сейшельские Острова, Сингапур, Словакия, Словения, Сомали, Южная Африка, Испания, Шри-Ланка, Судан , Суринам, Швеция, Швейцария, Тайвань, Танзания, Таиланд, Того, Тринидад и Тобаго, Тунис, Турция, Уганда, Украина, Объединенные Арабские Эмираты, Великобритания, Уругвай, США, Узбекистан, Венесуэла, Вьетнам, Замбия, Зимбабве, Доминика, Гренада, Мозамбик, Сирия, Тимор-Лешти, Белиз, Лаос, Ливия, Западный берег и сектор Газа, Гвинея-Бисау, Мали, Сент-Китс и Невис, Косово, Бирма, М.С. Зандам, Ботсвана, Бурунди, Сьерра-

Леоне, Малави, Южный Судан, Западная Сахара, Сан-Томе и Принсипи, Йемен.

3.2 Описание программного кода

Приведем описание основных блоков программного кода. В приложении используются следующие компоненты:

- библиотека numpy, реализующая наиболее распространённые математические функции;
- библиотека matplotlib, отвечающая за визуальную составляющую приложения в части построения графиков;
- библиотека pandas, которая используется импорта-экспорта и преобразования статистических данных;
- библиотека datetime, отвечающая за работу с датами;
- библиотека scipy, из которой используются методы для оптимизации параметров логистической функции.

Блок программного кода, отвечающего за подключение библиотек, показан на рисунке 3.1

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import pandas as pd
from IPython.display import display
import scipy as sp
from datetime import datetime
from scipy.optimize import minimize
```

Рисунок 3.1 – Библиотеки, используемые в приложении

Загрузка статистических данных с информацией о ежедневном количестве смертей от COVID-19 осуществляется из csv-файла,

размещенного в репозитории на github. Для загрузки данных из файла используется стандартный метод read_csv, реализованный в библиотеке pandas. В качестве параметра в данном случае указывается разделитель, применяемый к разметке csv для отделения значений друг от друга (рисунок 3.2).

```
coron = pd.read_csv('time_series_covid19_deaths_global.csv',sep=",")
coron
```

Рисунок 3.2 – Загрузка статистических данных по смертям от COVID-19

Полученные из файла данные сохраняются в переменной coron типа dataframe (специальный тип переменной для хранения таблиц, реализуемый библиотекой pandas). Для возможности визуального контроля результатов получения данных из csv текущее содержимое переменной coron выводится на экран в виде таблицы так, как это показано на рисунке 3.3. Также под таблицей выводится количество строк и столбцов.

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20
0	NaN	Afghanistan	33.000000	65.000000	0	0	0	0
1	NaN	Albania	41.153300	20.168300	0	0	0	0
2	NaN	Algeria	28.033900	1.659600	0	0	0	0
3	NaN	Andorra	42.506300	1.521800	0	0	0	0
4	NaN	Angola	-11.202700	17.873900	0	0	0	0
...	...	...	...	...	...	...	...	...
259	Saint Pierre and Miquelon	France	46.885200	-56.315900	0	0	0	0
260	NaN	South Sudan	6.877000	31.307000	0	0	0	0
261	NaN	Western Sahara	24.215500	-12.885800	0	0	0	0
262	NaN	Sao Tome and Principe	0.186360	6.613081	0	0	0	0
263	NaN	Yemen	15.552727	48.516388	0	0	0	0

264 rows × 103 columns

Рисунок 3.3 – Структура исходных статистических данных по смертям от COVID-19

Структура таблицы, хранящейся в переменной coron, полностью повторяет таблицу в исходном файле со статистическими данными. Однако такая структура неудобна для анализа. Поэтому следующим этапом является преобразование структуры таблицы.

```
all_countrys=coron["Country/Region"].unique()
print(len(all_countrys))
coron.columns = coron.columns.str.replace('Country/Region','Date')
coron.set_index("Date", inplace=True)
del coron['Province/State']
del coron['Lat']
del coron['Long']
coron
```

```
185
```

Рисунок 3.4 – Удаление ненужных столбцов

Для этого производится удаление из таблицы не нужных столбцов:

- столбец «Province/State», который содержит названия регионов различных стран;
- столбцы «Lat» и «Long», которые содержат координаты широты и долготы центра региона.

Так как на следующем шаге планируется произвести транспонирование таблицы с данными (разворот структуры таблицы на 90 градусов), то предварительно производится переименование столбца «Country/Region» в «Date». После этого столбец «Date» объявляется новым индексом таблицы путем вызова стандартного метода set_index().

Промежуточный результат преобразования структуры данных переменной coron показан на рисунке 3.5. В качестве столбцов таблицы используются даты с шагом в 1 день, начиная с 22 января 2020 года. В качестве строк используются страны, название которых приведены в алфавитном порядке. На пересечении столбцов и строк находятся целые значения, указывающие на количества смертей, зарегистрированных в этот день.

Date	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20
Afghanistan	0	0	0	0	0	0	0	0
Albania	0	0	0	0	0	0	0	0
Algeria	0	0	0	0	0	0	0	0
Andorra	0	0	0	0	0	0	0	0
Angola	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
France	0	0	0	0	0	0	0	0
South Sudan	0	0	0	0	0	0	0	0
Western Sahara	0	0	0	0	0	0	0	0
Sao Tome and Principe	0	0	0	0	0	0	0	0
Yemen	0	0	0	0	0	0	0	0

264 rows × 99 columns

Рисунок 3.5 – Промежуточный результат корректировки данных

Следующим шагом производится транспонирование таблицы, содержащейся в переменной coron и преобразование столбца с датами из строкового типа в тип datetime (рисунок 3.6).

```
coron = coron.T
coron.index = pd.to_datetime(coron.index)
#coron
```

Рисунок 3.6 – Транспонирование данных

В исходных данных приведена более детализованная информация по смертям, разбитая не только по странам, но еще и по регионам. Это можно

увидеть, выполнив операцию среза, оставив столбцы с названием «China» (рисунок 3.7).

Каждый столбец соответствует определенному региону Китая.

```
coron["China"]
```

Date	China	China	China	China	China	China	China	China	China	China	...
2020-01-22	0	0	0	0	0	0	0	0	0	0	...
2020-01-23	0	0	0	0	0	0	0	0	0	1	...
2020-01-24	0	0	0	0	0	0	0	0	0	1	...
2020-01-25	0	0	0	0	0	0	0	0	0	1	...
2020-01-26	0	0	0	0	0	0	0	0	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...
2020-04-25	6	8	6	1	2	8	2	2	6	6	...
2020-04-26	6	9	6	1	2	8	2	2	6	6	...
2020-04-27	6	9	6	1	2	8	2	2	6	6	...
2020-04-28	6	9	6	1	2	8	2	2	6	6	...
2020-04-29	6	9	6	1	2	8	2	2	6	6	...

99 rows × 33 columns

Рисунок 3.7 – Вывод данных о смертях в различных регионах Китая

Так как мы планируем анализировать данные в целом по странам, то необходимо выполнить процедуру консолидации данных по странам. В программном коде это реализовано за счет использования лямбда-функции, которая вызывается при группировке данных по названию столбцов (рисунок 3.8).

После консолидации при выполнении среза по столбцам «China» мы получаем единственный столбец с просуммированными значениями смертей по всем регионам (рисунок 3.8).

```
coron = coron.groupby(level=0, axis=1).apply(lambda x: x.apply(sum, axis=1))
coron["China"]
```

```
2020-01-22      17
2020-01-23      18
2020-01-24      26
2020-01-25      42
2020-01-26      56
               ...
2020-04-25    4636
2020-04-26    4637
2020-04-27    4637
2020-04-28    4637
2020-04-29    4637
Name: China, Length: 99, dtype: int64
```

Рисунок 3.8 – Группировка данных, результат группировки на примере Китая

Для того, чтобы знать список всех стран, доступных для анализа производится сохранение названий столбцов в список name_cols (рисунок 3.9).

```
name_cols = coron.columns.tolist()
```

Рисунок 3.9 – Сохранение массива с названиями всех стран, присутствующих в выборке данных

Для удобства прогнозирования и наглядности целесообразно ввести альтернативную временную шкалу, отсчитывающую номер дня от начала года. Для этого создается ряд данных типа series (рисунок 3.10).

```
X_data = []
for i in range(len(coron)):
    X_data.append(22 + i)
X_index = coron.index.values
```

```
X_Series = pd.Series(X_data, X_index)
coron.insert(0, "X", X_data)
#X_Series
```

Рисунок 3.10 – Формирование нового столбца X (номер дня от начала года)

Результатом подготовки данных для анализа является таблица, содержащаяся в переменной coron, которая демонстрируется пользователю на экране (рисунок 3.11).

```
coron
```

Date	X	Afghanistan	Albania	Algeria	Andorra	Angola	Antigua and Barbuda	Argentina
2020-01-22	22	0	0	0	0	0	0	0
2020-01-23	23	0	0	0	0	0	0	0
2020-01-24	24	0	0	0	0	0	0	0
2020-01-25	25	0	0	0	0	0	0	0
2020-01-26	26	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...
2020-04-25	116	47	27	419	40	2	3	185
2020-04-26	117	50	28	425	40	2	3	192
2020-04-27	118	57	28	432	40	2	3	197
2020-04-28	119	58	30	437	41	2	3	207
2020-04-29	120	60	30	444	42	2	3	214

99 rows × 187 columns

Рисунок 3.11 – Конечный результат подготовки данных для анализа

Следующим этапом после подготовки данных является их анализ с целью прогнозирования распространения COVID-19. Вообще приложение поддерживает анализ данных любого набора из 185 стран, но в нашем случае прогнозы рассчитываются для 11 крупнейших стран (Китай, Франция, Иран, Италий, Испания, Англия, США, Бельгия, Германия, Нидерланды, Швейцария), для всего мира целиком и также для любой выбранной пользователем страны.

Для этого в коде объявляются переменные, куда помещаются данные о количестве смертей от COVID-19 из соответствующих столбцов подготовленной таблицы (рисунок 3.12). Переменная anco зарезервирована, чтобы пользователь мог добавить любую страну, отсутствующую в списке 11 крупнейших стран.

```
X = corona['X']
chi = corona['China']
fr = corona['France']
ir = corona['Iran']
it = corona['Italy']
sp = corona['Spain']
uk = corona['United Kingdom']
us = corona['US']
bg = corona['Belgium']
gm = corona['Germany']
nt = corona['Netherlands']
sw = corona['Switzerland']
tot = corona['Total']

# название любой страны на англ. или Total
# для прогноза по всем странам:
any_county_name="Total"
anco = corona[any_county_name]
```

Рисунок 3.12 – Формирование серий данных на основе столбцов

```
dchi = chi[1:].values - chi[:-1].values
dfr = fr[1:].values - fr[:-1].values
dit = it[1:].values - it[:-1].values
diran = ir[1:].values - ir[:-1].values
dsp = sp[1:].values - sp[:-1].values
duk = uk[1:].values - uk[:-1].values
dus = us[1:].values - us[:-1].values
dbg = bg[1:].values - bg[:-1].values
dgm = gm[1:].values - gm[:-1].values
dnt = nt[1:].values - nt[:-1].values
dsw = sw[1:].values - sw[:-1].values
dtot = tot[1:].values - tot[:-1].values

danco =  anco[1:].values - anco[:-1].values
```

Рисунок 3.13 – Расчет новый данных: прирост смертей за день

Так как планируется анализировать не только изменение количество смертей с течением времени, но и изменение ежедневного прироста смертей, то необходимо рассчитать дополнительные данные на каждую дату – разницу между количеством смертей от COVID-19 сегодня и вчера. Для этого объявляются дополнительные переменные, причем переменная danco зарезервирована под любую страну, выбранную пользователем (рисунок 3.13).

Расчет разницы между количеством умерших на текущую дату и предшествующим днём в программном коде реализован с помощью операции среза. Выражение «[1:]» означает, что рассматриваются все элементы ряда кроме первого, а выражение «[:-1]» позволяет ссылаться все время на предыдущий элемент ряда (рисунок 3.13).

Следующем шагом в программном коде задаётся интервал времени, в котором статистические данные будут учтены. При этом на отрезке времени, где статистических данных нет, будет составлен прогноз. По умолчанию временной отрезок для аппроксимации и прогнозирования задан, как 180 дней, начиная с 20 января 2020 (рисунок 3.14).

```
X_long = np.arange(20, 200)
time_long = pd.date_range('2020-01-20', periods=180)
```

Рисунок 3.14 – Задание параметров для аппроксимации и прогнозирования

Аппроксимация статистических данных функцией заключается в подборе таких параметров A_0, A_1, A_2 логистической функции, которые обеспечивают наименьшую сумму квадратов ошибок. Ошибка – это разница на выбранную дату между тем значением, которое выдает функция, и фактическим значением, полученным из исходной таблицы. Рассчитывается по всем датам с известным количеством смертей от COVID-19.

Программный код, реализующий функцию resLogisticAnco, возвращающую сумму квадратов ошибок, представлен на рисунке 3.15. teor – это массив значений, возвращаемых логистической функций с параметрами A_0, A_1, A_2, anco – массив значений, полученных из таблицы со статистическими данными, x – номер дня, для которого рассчитывается ошибка.

Так как у каждой страны разный набор значений A_0, A_1, A_2, то функция для каждой отдельной страны задается отдельно (рисунок 3.15). Конкретно функция resLogisticAnco зарезервирована под страну, выбираемую для анализа пользователем.

```
def resLogisticAnco(coefficents):
    A0 = coefficents[0]
    A1 = coefficents[1]
    A2 = coefficents[2]
    teor = A0 / (1 + np.exp(A1 * (X.ravel() - A2)))

    return np.sum((teor - anco) ** 2)
```

Рисунок 3.15 – Задание функции, возвращающей сумму квадратов ошибок логистической функции при сравнении с фактическими данными

Код для запуска поиска оптимальных параметров A_0, A_1, A_2 показан на рисунке 3.16. minimize() является стандартным методом библиотеки SciPy, который решает задачу минимизации выбранной функции (в данном случае функции resLogisticAnco). В квадратных скобках задаются начальные значения параметров A_0, A_1, A_2, от которых будет осуществляться поиск оптимальных значений.

```
minim = minimize(resLogisticAnco, [16200, -.16, 80])
minim.x
```

```
array([ 2.55869179e+05, -1.18201737e-01,  1.04124904e+02])
```

Рисунок 3.16 – Поиск коэффициентов логистической функции, при которых сумма квадратов ошибок минимальна

Найденные значения параметров A_0, A_1, A_2 выводятся на экран пользователю.

Следующим этапом является визуализация аппроксимации статистических данных логистической функции, а также прогнозирование изменения количества смертей от COVID-19 в выбранной стране.

Для этого рассчитанные коэффициенты A_0, A_1, A_2 подставляются в функцию. С помощью библиотеки matplotlib рисуется декартовая система координат, где полученная функция отображается синей линией. При этом исходные данные отображаются на том же графике в виде черных крестов. Программный код для отображения фактических данных по числу смертей от COVID-19 и графика аппроксимирующей логистической функции для выбранной пользователем страны показан на рисунке 3.17.

```
plt.figure(figsize=(10,6))
teorAnco = minim.x[0] / (1 + np.exp(minim.x[1] * (X_long - minim.x[2])))
plt.plot(X_long[:120], teorAnco[:120],'b', label='Аппроксимация и прогноз')
plt.xticks(X_long[:120][::3], time_long.date[:120][::3], rotation='90');
tit = 'Количество умерших всего, '+ any_county_name
plt.title(tit, Size=12);
plt.plot(X,anco,'kx', label='Фактические данные')
plt.grid()
plt.legend()
plt.ylabel('Количество умерших человек')
```

```
Text(0, 0.5, 'Количество умерших человек')
```

Рисунок 3.17 – Задание параметров для отображения фактических данных по числу смертей от COVID-19 и графика аппроксимирующей логистической функции

Теперь, когда данные об изменении количества умерших от COVID-19 для выбранной пользователем страны визуализированы, проводится аналогичная визуализация изменения ежедневного прироста числа умерших (новых смертей в день).

На основе логистической функции и полученных ранее коэффициентов A_0, A_1, A_2 рассчитывается ряд значений как разница между количеством смертей в рассматриваемую дату и количеством смертей за день до рассматриваемой даты (берутся значения выдаваемые функцией). Данный ряд значений отображается на графике синей линией. Такой же ряд, но рассчитанный по фактическим значениям, отображается на графике в виде черных крестов.

Программный код для отображения данных по дневному приросту смертей от COVID-19 в виде графика для выбранной пользователем страны показан на рисунке 3.18.

```
plt.figure(figsize = (10,6))
plt.grid()
tit = 'Количество новых смертей в день, '+ any_county_name
plt.title(tit, Size=12);
plt.plot(X[1:], danco,'kx', label='Фактические данные')
plt.xticks(X_long[1:120][::3], time_long.date[1:120][::3], rotation='90');
plt.plot(X_long[1:120], teorAnco[1:120] - teorAnco[:119],'b', label='Аппрок
plt.ylabel('Количество умерших человек')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x2aace208>
```

Рисунок 3.18 –Задание параметров для отображения данных по дневному приросту смертей от COVID-19

Помимо анализа и составления прогноза для выбранной пользователем страны в приложении реализован анализ 11 крупнейших стран (их перечень приводился выше). Примеры кода для анализа данных США представлены на рисунках 3.19 – 3.21.

```
def resLogisticUs(coefficents):
    A0 = coefficents[0]
    A1 = coefficents[1]
    A2 = coefficents[2]
    teor = A0 / (1 + np.exp(A1 * (X.ravel() - A2)))
    return np.sum((teor - us) ** 2)
```

```
minim = minimize(resLogisticUs, [3200, -.16, 100])
minim.x
```

```
array([ 6.56096656e+04, -1.50549256e-01,  1.06074867e+02])
```

Рисунок 3.19 – Функция для поиска оптимальных коэффициентов логистической функции по данным США

```
plt.figure(figsize=(10,6))
teorUS = minim.x[0] / (1 + np.exp(minim.x[1] * (X_long - minim.x[2])))
plt.plot(X_long[:120], teorUS[:120],'b', label='Аппроксимация и прогноз')
plt.xticks(X_long[:120][::3], time_long.date[:120][::3], rotation='90');
plt.title('Количество умерших всего, США', Size=12);
plt.plot(X,us,'kx', label='Фактические данные')
plt.grid()
plt.legend()
plt.ylabel('Количество умерших человек')
```

```
Text(0, 0.5, 'Количество умерших человек')
```

Рисунок 3.20 – Задание параметров для отображения данных по дневному приросту смертей от COVID-19 в США

```
plt.figure(figsize = (10,6))
plt.grid()
plt.title('Количество новых смертей в день, США', Size=12);
plt.plot(X[1:], dus, 'kx', label='Фактические данные')
#plt.plot(X[1:], dus, 'ro')
plt.xticks(X_long[1:120][::3], time_long.date[1:120][::3], rotation='90');
plt.plot(X_long[1:120], teorUS[1:120] - teorUS[:119],'b', label='Аппроксима
plt.ylabel('Количество умерших человек')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x3a6e5448>
```

Рисунок 3.21 – Задание параметров для отображения данных по дневному приросту смертей от COVID-19 в США

В приложении также существует возможность получения прогноза по числу смертей от COVID-19 на конкретную дату. Для этого задается номер дня от начала года, для которого требуется получить прогноз. Данное значение хранится в переменной alone_time. Результат преобразования данного значения в конкретную дату для удобства выводится на экран. Программный код представлен на рисунке 3.22.

```
# Прогноз на определенную дату
alone_time = 110
print('Прогнозы на:',time_long.date[alone_time])
```

```
Прогнозы на: 2020-05-09
```

Рисунок 3.22 – Задание дня прогнозирования (номер дня от начала года)

Затем для заданного дня выводится рассчитанное с использованием логистической функции значение, округленное до целого числа. Так как в качестве имени страны пользователем было задано значение «Total», то

будет выдан мировой прогноз на заданную дату. Программный код и отображаемый результат представлены на рисунке 3.23.

```
print(any_county_name)
alone = teorAnco[:120]
d_alone = teorAnco[1:120] - teorAnco[:119]
print('Общее число смертей:',round(alone[alone_time]))
print('Прирост смертей за последние сутки:',round(d_alone[alone_time]))
```

```
Total
Общее число смертей: 244393.0
Прирост смертей за последние сутки: 1228.0
```

Рисунок 3.23 – Вывод прогноза на заданную дату

Для 11 крупных стран прогноз на эту же дату генерируется автоматически без участи пользователя. Для примера ниже приведен код, отвечающий за прогноз по COVID-19 в США (рисунок 3.24).

```
print("США")
alone = teorUS[:120]
d_alone = teorUS[1:120] - teorUS[:119]
print('Общее число смертей:',round(alone[alone_time]))
print('Прирост смертей за последние сутки:',round(d_alone[alone_time]))
```

```
США
Общее число смертей: 63868.0
Прирост смертей за последние сутки: 238.0
```

Рисунок 3.24 – Вывод прогноза на заданную дату для США

Все представленные во втором разделе данные (графики, аппроксимирующие зависимости) получены с помощью описанного выше программного обеспечения.

Выводы по разделу

На языке Python было разработано программное обеспечение, позволяющее получать актуальные данные о количестве зарегистрированных смертей по вирусу COVID-19 и пересчитывать параметры логистических функций с учетом новых данных. Результат аппроксимации исходных данных представляется визуально в виде графиков: количество зарегистрированных смертей в зависимости от даты, прирост количества зарегистрированных смертей в зависимости от даты.

ЗАКЛЮЧЕНИЕ

По результатам собранных материалов, проведенных теоретических и практических исследований можно сформулировать следующие основные выводы.

1. Актуальность темы обусловлена заявлением Всемирной организацией здравоохранения о признании COVID-19 в 2020 году пандемией. Поэтому все исследования, направленные на изучение особенностей распространения COVID-19, являются в настоящее время актуальными.

2. Обзор литературных источников показал, что современные работы в области анализа данных направленны на исследования способов ранней диагностики COVID-19, определение групп риска на основе анализа статистики, прогнозирования тяжести протекания заболевания, а также определения дат снижения прироста числа заболевших.

3. В работе предложен алгоритм прогнозирования распространения вирусной инфекции COVID-19, который заключается в выполнении следующих шагов. Загрузка статистических данных Всемирной организации здравоохранения о ежедневном количестве зарегистрированных смертей от вируса (отдельно по каждой стране). Затем аппроксимация полученных данных логистической функцией путем подбора параметров функции методом наименьших квадратов. Так как динамика распространения COVID-19 внутри каждой страны уникальна, то и параметры логистической функции для каждой страны свои собственные. Затем полученные функции используются для оценки изменения количества смертей по каждой стране в ближайшее время.

4. Результаты исследований точности прогнозирования с использованием предложенных подходов показали следующее. Получаемые прогнозы на основе логистических функций обладают приемлемой

точностью в краткосрочной перспективе (до 1 недели, точность 85%). С увеличением временного интервала от даты последнего получения фактических данных по COVID-19 точность прогнозирования уменьшается.

5. В рамках данного исследования на языке Python с использованием библиотек pandas, numpy, scipy, matplotlib, datetime было разработано программное обеспечение, позволяющее получать актуальные данные о количестве зарегистрированных смертей по вирусу COVID-19 и пересчитывать с учетом новых данных параметры логистических функций. Результат аппроксимации исходных данных представляется графически в виде графиков временных рядов: количество зарегистрированных смертей в зависимости от даты и прирост количества зарегистрированных смертей в зависимости от даты. При этом в программном обеспечении есть возможность получения прогноза на указанную дату.

Таким образом, все поставленные задачи были выполнены и поставленная цель бакалаврской работы достигнута.

СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

1. Аверкин, А.Н. Гибридный подход для прогнозирования временных рядов на основании нейросети ANFIS и нечетких когнитивных карт / А.Н. Аверкин, С.А. Ярушев // Международная конференция по мягким вычислениям и измерениям. – Санкт-Петербург : Издатель Санкт-Петербургский государственный электротехнический университет "ЛЭТИ" им. В.И. Ульянова (Ленина) (Санкт-Петербург), 2017. – С. 467-470. – Текст: непосредственный.

2. Басинский, В.М. Алгоритм муравьиной колонии при решении задачи классификации и использование генетического алгоритма для подбора его параметров / В.М. Басинский, Ю.Г. Степин // Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2018) Электронный сборник статей I международной научно-практической конференции, посвященной 50-летию Полоцкого государственного университета. 2018. – Новополоцк : Издатель Учреждение образования «Полоцкий государственный университет», 2018. – С. 118-122. – Текст : непосредственный.

3. Буланов, О. Тестирование гипотезы о многообразии / О. Буланов, Ю. Янович // Информационные технологии и системы 2017 (ИТИС 2017). Уфа, 14-17 сентября 2017 г. – Уфа : Издатель Институт проблем передачи информации им. А.А. Харкевича РАН, Москва, 2017. – Р. 41 – 48. – Текст : непосредственный.

4. Вельдяйкин, Н. Алгоритм laplacian eigenmaps для точек вне обучающей выборки / Н. Вельдяйкин, Ю. Янович // Информационные технологии и системы 2017 (ИТИС 2017)Уфа, 14-17 сентября 2017 г. – Уфа : Издатель Институт проблем передачи информации им. А.А. Харкевича РАН, Москва, 2017. – Р. 74 – 80. – Текст : непосредственный.

5. Дорофеюк, А.А. Методология структурно-классификационного исследования сложно организованной информации в задачах интеллектуального анализа данных / А.А. Дорофеюк, А.Ю. Дорофеюк, // XII всероссийское совещание по проблемам управления ВСПУ-2014. Москва, 16-19 июля 2014 г. – Москва : Издатель Институт проблем управления им. В.А. Трапезникова РАН Москва, 2014. – С. 8369 – 8381. – Текст : непосредственный.

6. Ильина, М.А. Универсальный алгоритм визуализации решений задачи классификации / М.А. Ильина, А.А. Тузовский // Современные проблемы естественных и технических наук. Новосибирск, 24-25 мая 2016 г. – Новосибирск : Издатель Новосибирский государственный архитектурно-строительный университет Сибстрин, Новосибирск, 2016. – С. 46 – 50. – Текст : непосредственный.

7. Рыцарев, И.А. Применение метода главных компонент для выявления семантических различий и анализа изменения положения в пространстве при анализе информационного контента сетевых сообществ / И.А. Рыцарев, Р.А. Парингер, А.В. Куприянов // V международная конференция и молодежная школа "информационные технологии и нанотехнологии". Самара, 21-24 мая 2019 г. – Самара : Издатель Новая техника , Самара, 2019. – С. 780 – 787. – Текст : непосредственный.

8. Сидорова, В.А. Выбор размерности и детальности данных дистанционного зондирования земли при кластеризации гистограммным иерархическим алгоритмом / В.А. Сидорова // Актуальные проблемы вычислительной и прикладной математики, Новосибирск, 19-23 октября 2015 г.– Новосибирск : Издатель Институт вычислительной математики и математической геофизики СО РАН, Новосибирск, 2015. – Р. 664 – 669. – Текст : непосредственный.

9. Чульдум, А.Ф. Пример создания C# .NET приложения для прогнозирования временного ряда с использованием адаптивной нейро-нечеткой системы вывода-ANFIS / А.Ф. Чульдум, У.А. Чульдум //

Информатизация образования: история, проблемы и перспективы сборник материалов Всероссийской научно-практической конференции, посвященной 70-летию со дня рождения первого ректора Тувинского государственного университета О.Б. Бузур-оола. 2016. – Кызыл : Издатель: Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования "Тувинский государственный университет", 2016. – С. 38-41. – Текст : непосредственный.

10. Agarwal, I. TensorFlow for Doctors / Isha Agarwal, Rajkumar Kolakaluri, Michael Dorin, Mario Chong // Annual International Symposium on Information Management and Big Data SIMBig 2019: 6th International Conference, SIMBig 2019, Lima, Peru, August 21–23, 2019, Proceedings. – Springer Nature Switzerland AG, 2020 – pp. 76-88. – Text : direct.

11. Cho, Y.-J. Rule Generation Using NN and GA for SARS-CoV Cleavage Site Prediction / Yeon-Jin Cho, Hyeoncheol Kim // International Conference on Knowledge-Based and Intelligent Information and Engineering Systems KES 2005 - Knowledge-Based Intelligent Information and Engineering Systems: 9th International Conference, Melbourne, Australia, September 14-16, 2005, Proceedings, Part III. – Springer-Verlag Berlin Heidelberg, 2005. – pp. 785-791. – Text : direct.

12. Cho, Y.-J. Cleavage Site Analysis Using Rule Extraction from Neural Networks / Yeun-Jin Cho, Hyeoncheol Kim // International Conference on Natural Computation ICNC 2005 - Advances in Natural Computation: First International Conference, Changsha, China, August 27-29, 2005, Proceedings, Part I. – Springer-Verlag Berlin Heidelberg, 2005 – pp. 1002-1008. – Text : direct.

13. John, M. Shiny Framework Based Visualization and Analytics Tool for Middle East Respiratory Syndrome / Maya John, Hadil Shaiba // International Conference on Computing ICC 2019 - Advances in Data Science, Cyber Security and IT Applications: First International Conference on Computing, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part I. – Springer Nature Switzerland AG, 2019. – pp. 193-202. – Text : direct.

14. Li, X. Method for Recognition Pneumonia Based on Convolutional Neural Network / L. Xin, D. Gao, H. Hao // International Conference of Pioneering Computer Scientists, Engineers and Educators: 5th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2019, Guilin, China, September 20–23, 2019, Proceedings, Part II. – Singapore: Springer Nature Singapore Pte Ltd., 2019. – P. 142-156. – Text : direct.

15. Li, J. Classification and Characteristics of TCM Syndromes of Chronic Respiratory Failure Based on Self-adaptive Fuzzy Inference System / Jiansheng Li, Haifeng Wang, Jinliang Hu, Jiehua Wang, Suyun Li, Minghang Wang, Ya Li // International Conference on Intelligent Computing ICIC 2010 – Advanced Intelligent Computing Theories and Applications: 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18-21, 2010. Proceedings. – Springer-Verlag Berlin Heidelberg, 2010. – pp. 266-272. – Text : direct.

16. Lidayova, K. Airway-Tree Segmentation in Subjects with Acute Respiratory Distress Syndrome / Kristina Lidayova, Duvan Alberto, Gomez Betancur, Hans Frimmel, Marcela Hernandez, Hoyos Maciej, Orkisz Orjan // Scandinavian Conference on Image Analysis SCIA 2017 - Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromso, Norway, June 12–14, 2017, Proceedings, Part II. – Springer International Publishing AG, 2017. – pp. 76-87. – Text : direct.

17. Pfannschmidt, K. Evaluating Tests in Medical Diagnosis: Combining Machine Learning with Game-Theoretical Concepts / Karlson Pfannschmidt, Eyke Hullermeier, Susanne Held, Reto Neiger // International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU 2016 - Information Processing and Management of Uncertainty in Knowledge-Based Systems: 16th International Conference, IPMU 2016, Eindhoven, The Netherlands, June 20-24, 2016, Proceedings, Part I. – Springer International Publishing Switzerland, 2016. – pp. 450-461. – Text : direct.

18. Robust Association of Pathological Respiratory Events in SAHS Patients: A Step towards Mining Polysomnograms / Abraham Otero, Paulo Felix // International Work-Conference on Artificial Neural Networks IWANN 2009 - Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living: 10th International Work-Conference on Artificial Neural Networks, Salamanca, Spain, June 10-12, 2009. Proceedings, Part II – Springer-Verlag Berlin Heidelberg, 2009. – pp. 1020-1027. – Text : direct.

19. Sguanci, L. Modeling Evolutionary Dynamics of HIV Infection / Luca Sguanci, Pietro Lio, Franco Bagnoli // International Conference on Computational Methods in Systems Biology CMSB 2006: Computational Methods in Systems Biology, International Conference, Trento, Italy, October 18-19, 2006. Proceedings. – Springer-Verlag Berlin Heidelberg, 2006 – pp. 196-211. – Text : direct.

20. Teng, H. Applying of Adaptive Threshold Non-maximum Suppression to Pneumonia Detection / Hao Teng, Huijuan Lu, Minchao Ye, Ke Yan, Zhigang Gao, Qun Jin // International Conference on Bio-Inspired Computing BIC-TA 2019: Bio-inspired Computing: Theories and Applications, 14th International Conference, BIC-TA 2019, Zhengzhou, China, November 22–25, 2019, Revised Selected Papers, Part II. - Springer Nature Singapore Pte Ltd., 2020 – pp. 518-528. – Text : direct.